Nouvelles Méthodes d'Apprentissage Profond pour la Détection et la Reconnaissance d'Outils et d'Activités Chirurgicaux dans les Vidéos Laparoscopiques

Le Candidat:	Chinedu Innocent Nwoye
Directeur de thèse:	Prof. Nicolas Padoy (PhD)
Co-directeur de thèse:	Prof. Didier Mutter (MD)
Laboratoire:	CAMMA, AVR, ICube Laboratory
École Doctorale:	MSII, Université de Strasbourg, France

1 Introduction

Les technologies émergentes du bloc opératoire ont encouragé la conception de nombreux algorithmes d'apprentissage profond pour l'analyse automatisée du flux de travail chirurgical dans le but de fournir une intervention assistée par ordinateur (IAO) dans la salle d'opération [5]. Un ingrédient clé pour développer des systèmes d'IAO qui peuvent fournir une aide contextuelle à la décision en chirurgie laparoscopique est d'avoir une connaissance en temps réel de la présence des instruments chirurgicaux, de leur pose par rapport à la caméra, de l'anatomie sous-jacente, de leur trajectoire, et de comprendre leurs interactions avec les tissus environnants.

Dans la première partie de la thèse [1], nous étudions et construisons des modèles capables d'exploiter des données faiblement annotées pour la détection, la localisation et le suivi des instruments chirurgicaux. La détection et le suivi des instruments faciliteront la compréhension des interactions outil-tissu, fourniront des signaux sur les situations chirurgicales, aideront à la décision chirurgicale et à l'évaluation des compétences. Les travaux existants sur la détection d'instruments chirurgicaux reposent sur une supervision complète : une situation dans laquelle les modèles de détection et de suivi sont entraînés sur des données pour lesquelles les positions spatiales des instruments sont annotées manuellement. Cependant, la création d'annotations spatiales telles que des régions rectangulaires ou des masques de pixels est coûteuse et fastidieuse. En outre, la plupart des jeux de données disponibles dans le domaine ne contiennent que des étiquettes de présence binaires qui sont générées par un simple marquage de 0 ou 1 pour indiquer la présence ou l'absence d'instruments chirurgicaux. Puisqu'il est plus facile de générer ces étiquettes binaires, nous avons proposé de les exploiter pour une tâche spatiale beaucoup plus complexe telle que la localisation d'instruments chirurgicaux. Notre modèle proposé exploite également les informations temporelles inhérentes aux données vidéo pour apprendre à suivre les instruments.

Dans la deuxième partie de la thèse [2], nous nous appuyons sur les recherches les plus avancées pour reconnaître les activités des instruments chirurgicaux. Les travaux existants sur la reconnaissance d'activité se concentrent principalement sur la reconnaissance de phase [3, 11], d'étape [10], d'événement [8], de geste [4] ou même d'action définie par un verbe [7]. À ces niveaux de granularité, les modèles de reconnaissance laissent de côté une sémantique essentielle à l'IA: les interactions avec l'anatomie. Par conséquent, nous proposons une analyse plus détaillée de la reconnaissance d'activités fines représentant les interactions instrument-tissu dans les vidéos endoscopiques. Nous modélisons ces activités sous forme de triplets d'action chirurgicale ⟨instrument, verbe, cible⟩ et développons des modèles d'apprentissage profond pour reconnaître ces triplets. Le triplet représente l'instrument utilisé, l'action effectuée et l'anatomie sur laquelle on agit, comme défini dans l'ontologie existante

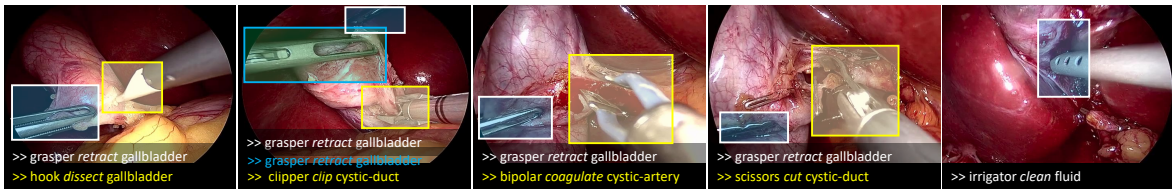

Figure 1: Échantillons d'images chirurgicales montrant des étiquettes d'instance de triplet d'action.

[9, 6]. Leur reconnaissance fournit des informations plus détaillées sur la situation chirurgicale, ce qui aide à mieux comprendre l'interaction outil-tissu pendant la chirurgie et ajoute des informations cruciales nécessaires à une IA sûre et efficace. La reconnaissance de triplets peut par exemple être utile en peropératoire pour la surveillance des points de contrôle de sécurité, l'anticipation des actions pour une intervention précoce et en postopératoire pour le sous-titrage vidéo, la génération de rapports postopératoires et l'indexation ainsi que la récupération de vidéos spécifiques à l'action pour former les chirurgiens. Pour soutenir la recherche dans cette direction, avec l'aide de nos collaborateurs cliniques, nous avons généré de nouveaux ensembles de données à grain fin (*CholecT40* et *CholecT50*), les premiers du genre, pour la reconnaissance de triplets d'action chirurgicale. Voici quelques exemples de triplets d'action chirurgicale:⟨grasper, retract, gallbladder⟩, ⟨hook, dissect, omentum⟩, ⟨scissors, cut,cystic-duct⟩, également illustrés dans la figure 1.

2 Tâches et Méthodes

2.1 Méthode Faiblement Supervisée pour la Localisation et le Suivi des Outils

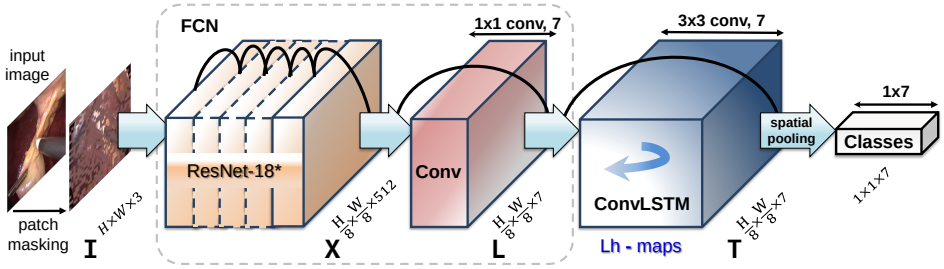

Figure 2: L'architecture du Traqueur ConvLSTM proposé (la configuration de la variante XLT).

Le modèle proposé tel qu'illustré à la figure 3 est une jointure de réseaux de neurones convolutifs (CNN) + mémoire convolutive à long court terme (ConvLSTM) qui est entraînée de bout en bout en tant que réseau entièrement convolutif. Notre première contribution principale ici est de proposer une modélisation faiblement supervisée de la localisation spatiale à l'aide d'une couche de convolution à 7 canaux qui sert de cartes thermiques de localisation (Lh-maps), également appelées cartes d'activation de classe (CAM). Notre deuxième contribution principale utilise un ConvLSTM faiblement supervisé pour lisser le pic d'activation dans les cartes Lh en tirant parti de la cohérence spatio-temporelle à travers les images vidéo pour apprendre des trajectoires fluides pour les instruments.

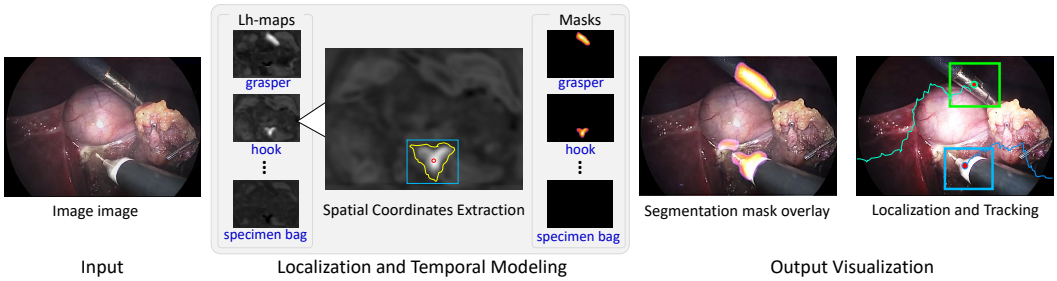

Figure 3: Séquence de traitement du modèle ConvLSTM.

2.2 Reconnaissance des Triplets d'Action Chirurgicale

Pour la deuxième partie, nous construisons progressivement 3 modèles d'apprentissage profond pour reconnaître les triplets d'action chirurgicale directement à partir de vidéos chirurgicales. Le premier modèle est connu sous le nom de *Tripnet* qui est un modèle centré sur l'instrument avec des branches d'apprentissage multi-tâches pour modéliser les différentes composantes du triplet, à savoir: instrument, verbe et cible. Le Tripnet s'appuie sur 2 nouveaux composants : le guide d'activation de classe (CAG) et les modules d'espace d'interaction tridimensionnel (3Dis). Le module CAG utilise

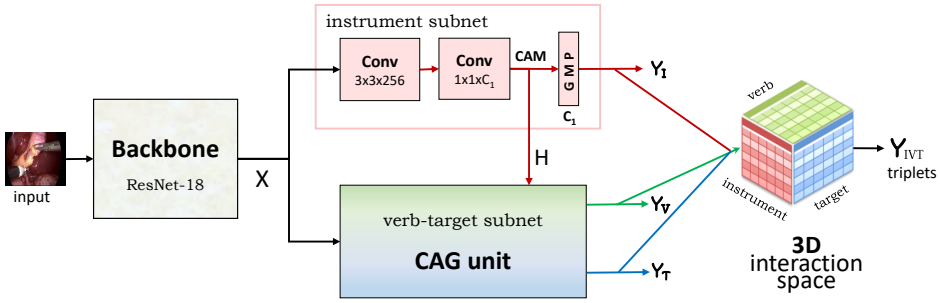

Figure 4: Tripnet : l'architecture du modèle proposé pour la reconnaissance de triplet d'action.

le repère d'apparence des instruments (CAM) pour conditionner l'espace de recherche du modèle pour la détection du verbe et de la cible, comme le montre la figure 4, alors que le 3Dis entraînable gère l'association tripartite de la relation des composants du triplet pour correctement reconnaître l'interaction. Le deuxième modèle, appelé Attention Tripnet, étend le Tripnet avec un mécanisme d'attention guidée par CAM (CAGAM) pour un conditionnement spatial plus précis et une meilleure détection du verbe et des composants cibles du triplet. Le dernier modèle est un modèle inspiré du Transformer, connu sous le nom de Rendezvous, qui tire parti de l'attention à plus long terme pour apprendre l'interaction outil-tissu. La nouveauté réside dans son utilisation de mécanismes multi-têtes à la fois d'auto-attention et d'attention croisée pour capturer l'interaction des trois composants d'un triplet. Le Rendez-vous est mieux décrit comme la réunion de plusieurs composants d'attention pour comprendre leur relation en tant que triplet.

3 Résultats et Conclusion

Table 1: Résultats quantitatifs sur la détection, la localisation et le suivi des outils.

Modèle	Détection	Localisation	Tracking: $\Theta = \mu\ (0.3 - 0.7)$	
	(% mAP)	(% $IoU \geq 0.5$)	MOTP	MOTA
Référence	87,7	21,0	62,1	23,9
Le nôtre	92,9	38,2	67,4	36,5

Cette méthode de suivi et de détection d'instruments proposée dans ce travail est validée sur le célèbre jeu de données Cholec80 pour la détection de présence, la localisation spatiale et le suivi de mouvement des instruments chirurgicaux dans les vidéos laparoscopiques où elle surpasse toutes les références comparées comme présenté dans le tableau 1 établissant ainsi une nouvelle performance à la pointe de la technologie.

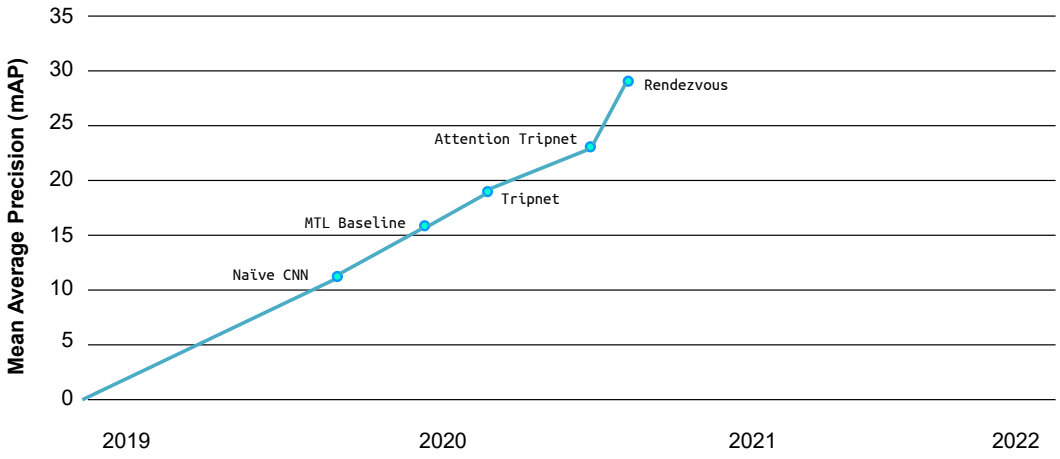

Figure 5: Résumé graphique des performances de tous les modèles évalués sur la reconnaissance de triplet d'action.

De plus, la figure 5 montre l'amélioration progressive du modèle proposé par rapport aux références pour la reconnaissance de triplets d'action chirurgicale, établissant ainsi une référence pour de futures analyses dans cette nouvelle tâche dans le domaine de la science des données chirurgicales.

Les modèles résultant de nos expériences possèdent le potentiel d'être déployés dans les systèmes IAO au bloc opératoire pour assister la chirurgie, ainsi que d'être utilisés dans le développement d'applications médicales.

Publications

[1] C. I. Nwoye, D. Mutter, J. Marescaux, and N. Padoy, "Weakly supervised convolutional lstm approach for tool tracking in laparoscopic videos," *International journal of computer assisted radiology and surgery.*, vol. 14, no. 6, pp. 1059–1067 (2019), **A remporté le prix IPCAI Audience Choice Award pour la meilleure présentation papier**.

[2] C. I. Nwoye, C. Gonzalez, T. Yu, P. Mascagni, D. Mutter, J. Marescaux, and N. Padoy, "Recognition of instrument-tissue interactions in endoscopic videos via action triplets," in *MICCAI*, 2020.

References

[3] Blum, T., Feußner, H., Navab, N.: Modeling and segmentation of surgical workflow from laparoscopic video. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 400–407 (2010)

[4] DiPietro, R., Lea, C., Malpani, A., Ahmidi, N., Vedula, S.S., Lee, G.I., Lee, M.R., Hager, G.D.: Recognizing surgical activities with recurrent neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 551–558 (2016)

[5] Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D.I., Wang, G., Eaton-Rosen, Z., Gray, R., Doel, T., Hu, Y., et al.: Niftynet: a deep-learning platform for medical imaging. Computer methods and programs in biomedicine **158**, 113–122 (2018)

[6] Katić, D., Wekerle, A.L., Gärtner, F., Kenngott, H., Müller-Stich, B.P., Dillmann, R., Speidel, S.: Knowledge-driven formalization of laparoscopic surgeries for rule-based intraoperative context-aware assistance. In: International Conference on Information Processing in Computer-Assisted Interventions. pp. 158–167. Springer (2014)

[7] Khatibi, T., Dezyani, P.: Proposing novel methods for gynecologic surgical action recognition on laparoscopic videos. Multimedia Tools and Applications **79**(41), 30111–30133 (2020)

[8] Loukas, C., Georgiou, E.: Smoke detection in endoscopic surgery videos: a first step towards retrieval of semantic events. The International Journal of Medical Robotics and Computer Assisted Surgery **11**(1), 80–94 (2015)

[9] Neumuth, T., Strauß, G., Meixensberger, J., Lemke, H.U., Burgert, O.: Acquisition of process descriptions from surgical interventions. In: International Conference on Database and Expert Systems Applications. pp. 602–611 (2006)

[10] Ramesh, S., Dall'Alba, D., Gonzalez, C., Yu, T., Mascagni, P., Mutter, D., Marescaux, J., Fiorini, P., Padoy, N.: Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures. arXiv preprint arXiv:2102.12218 (2021)

[11] Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging **36**(1), 86–97 (2016)